DECLARATION

I, Kentaro OGAWA, who undersigned below, declare that I have translated the priority document for Japanese Patent Application No. 2003-383072 and that the attached English document is a true and correct translation for said priority document to the best of my knowledge and belief.

Signature of Translator: _____

Kentaro OGAWA

Date: Jan. 5, 2011

[Name of Document] Patent Application
[Reference Number] H103234301
[Date of submission] November 12, 2003
[Addressed to] Commissioner of Patent Office
[International Classification] G10L 15/00
[Inventor]
[Domicile] c/o HONDA R&D CO., LTD., 4-1, Chuo 1-chome, Wako-shi, Saitama, Japan
[Name] Kazuhiro NAKADAI
[Inventor]
[Domicile] c/o HONDA R&D CO., LTD., 4-1, Chuo 1-chome, Wako-shi, Saitama, Japan
[Name] Hiroshi OKUNO
[Inventor]
[Domicile] c/o HONDA R&D CO., LTD., 4-1, Chuo 1-chome, Wako-shi, Saitama, Japan
[Name] Hiroshi TSUJINO
[Applicant]
[Identification Number] 000005326
[Name] HONDA MOTOR CO., LTD.
[Representative]
[Identification Number] 100064414
[Attorney]
[Name] Michizo ISONO
[Phone Number] 03-5211-2488
[Indication of Official Fee]
[Number of Payment Ledger] 015392
[Amount] 21,000 Yen
[Attached Documents]
[Name of Document] Claims 1
[Name of Document] Specification 1
[Name of Document] Drawing 1
[Name of Document] Abstract 1
[Number of General Power of Attorney] 9713945

[WHAT IS CLAIMED IS]

[Claim 1] An automatic speech recognition system, which recognizes speeches of a specified speaker from acoustic signals detected by a plurality of microphones and converts the speeches into character information, the system comprising: a sound source localization module which localizes a sound direction corresponding to the specified speaker based on the acoustic signals detected by the plurality of microphones; a sound source separation module which separates speech signals of the specified speaker from the acoustic signals based on the sound direction localized by the sound source localization module; a feature extractor which extracts features of speech signals contained in the acoustic signals based on the speech signals separated by the sound source separation module; an acoustic model memory which stores direction-dependent acoustic models that correspond to a plurality of directions at intervals; an acoustic model composition module which composes an acoustic model adjusted to the sound direction, which is localized by the sound source localization module, based on the direction-dependent acoustic models in the acoustic model memory, the acoustic model composition module storing the acoustic model in the acoustic model memory; and a speech recognition module which recognizes the features extracted by the feature extractor and converts the features into character information by using the acoustic model composed by the acoustic model composition module.

[Claim 2] A system according to claim 1, wherein the sound source localization module is configured to execute a process comprising: performing a frequency analysis for the acoustic signals detected by the microphones to extract harmonic structures; acquiring an intensity difference and a phase difference for the harmonic structures extracted through the plurality of microphones; acquiring belief factors for

a sound direction based on the intensity difference and the phase difference, respectively; and determining a most probable sound direction.

[Claim 3] A system according to any one of claim 1 or 2, wherein the sound source separation module employs an active direction-pass filter so as to separate speeches, the filter being configured to execute a process comprising: separating speeches of a narrower directional band when a sound direction, which is localized by the sound source localization module, lies close to a front, which is defined by an arrangement of the plurality of microphones; and separating speeches of a wider directional band when the sound direction lies apart from the front.

[Claim 4] A system according to any one of claims 1 to 3, wherein the acoustic model composition module is configured to compose an acoustic model for the sound direction by applying weighted linear summation to the direction-dependent acoustic models in the acoustic model memory, and weights introduced into the linear summation are determined by training.

[Claim 5] A system according to any one of claims 1 to 4, further comprising a speaker identification module for identifying the speaker, wherein the acoustic model memory possesses the direction-dependent acoustic models for respective speakers, and wherein the acoustic model composition module is configured to execute a process comprising: based on direction-dependent acoustic models of a speaker who is identified by the speaker identifying module and a sound direction localized by the sound source localization module, obtaining an acoustic model for the sound direction based on the direction-dependent acoustic models in the acoustic model memory; and storing the acoustic model in the acoustic model memory.

SPECIFICATION

[Title of the Invention] SPEECH RECOGNITION DEVICE

[Detailed Description of the Invention]

[Field of Invention]

[0001]

The present invention relates to an automatic speech recognition system and, more particularly, to an automatic speech recognition system which is able to recognize speeches with high accuracy, even when a speaker and a moving object having the automatic speech recognition system are moving around.

[Prior Art]

[0002]

A technique for speech recognition, which has been recently developed so much as to reach practical use, has been started to apply to an area such as inputting of information in the form of speech. Also, research and development of robots has been flourishing, which induces a situation in which the technique for speech recognition technically plays a key role in putting a robot to practical use. This is ascribed to the fact that intelligently social interaction between a robot and a human requires the robots to understand human language, increasing the importance of accuracy achieved in speech recognition.

[0003]

There are several problems in conducting communication with a speaker, different from speech recognition, which is carried out in a laboratory by inputting speeches through a microphone which is placed near a mouth of the speaker.

For example, since there are various types of noise in an actual environment, it is not possible to succeed in speech recognition unless necessary speech signals are

separated from the noise. When there is a plurality of speakers, it is necessary to extract speeches of a specified speaker to be recognized. A Hidden Markov Model (HMM) is generally used for speech recognition. This model is not free of a problem that a recognition rate is adversely affected by the fact that a voice of a speaker sounds different according to positions of the speaker (direction relative to a microphone of an automatic speech recognition system).

[0004]

In view of the foregoing problems, a research group including the inventors of the present invention has disclosed a technique that performs localization, separation and recognition of a plurality of sound sources by active audition (see no-patent document 1).

This technique, which has two microphones provided at positions corresponding to ears of human, enables recognition of words uttered by one speaker when a plurality of speakers simultaneously utter words. More specifically speaking, the technique localizes the positions of the speakers based on acoustic signals entered through the two microphones and separates speeches for each speaker so as to recognize them. In this recognition, acoustic models of the each speaker are generated beforehand, which are adjusted to directions covering a range of -90 deg. to 90 deg. at intervals of 10 deg. as viewed from a moving object (such as a robot or the like, having an automatic speech recognition system). When speech recognition is performed, processes with using these acoustic models are carried out in parallel.

[No-patent document 1] : "A humanoid Listens to three simultaneous talkers by Integrating Active Audition and Face Recognition" Kazuhiro Nakadai, et al., IJCAI-03 Workshop on Issues in Designing Physical Agents for Dynamic Real-Time

Environments: World Modeling, Planning, Learning and Communicating, PP117-124.

[Means for Solving the Problems]

[0005]

However, the conventional technique described above has posed a problem that because a position of the speaker changes with respect to the moving object each time the speaker and the moving object relatively move, a recognition rate decreases if the speaker stands at a position, for which an acoustic model is not prepared in advance.

The present invention, which is created in view of the background described above, provides an automatic speech recognition system which is able to recognize with high accuracy while a speaker and a moving object are moving around.

[0006]

In order to solve the above problems, it is an object of the present invention to provide an automatic speech recognition system, which recognizes speeches of a specified speaker from acoustic signals detected by a plurality of microphones and converts the speeches into character information. The system comprises a sound source localization module which localizes a sound direction corresponding to the specified speaker based on the acoustic signals detected by the plurality of microphones; a sound source separation module which separates speech signals of the specified speaker from the acoustic signals based on the sound direction localized by the sound source localization module; a feature extractor which extracts features of speech signals contained in the acoustic signals based on the speech signals separated by the sound source separation module; an acoustic model memory which stores direction-dependent acoustic models that correspond to a plurality of

directions at intervals; an acoustic model composition module which composes an acoustic model adjusted to the sound direction, which is localized by the sound source localization module, based on the direction-dependent acoustic models in the acoustic model memory, the acoustic model composition module storing the acoustic model in the acoustic model memory; and a speech recognition module which recognizes the features extracted by the feature extractor and converts the features into character information by using the acoustic model composed by the acoustic model composition module.

[0007]

In the automatic speech recognition system described above, the sound source localization module localizes a sound direction, the sound source separation module separates only the speeches of the sound direction localized by the sound source localization module, the acoustic model composition module composes an acoustic model adjusted to a direction based on the sound direction and direction-dependent acoustic models and the speech recognition module performs speech recognition with the acoustic model.

Moreover, it may be preferable that the speech signals outputted by the sound source separation module are information including the speeches and include not only analogue speech signals themselves but also a digitized, an encoded signal and spectrum data obtained by frequency analysis.

[0008]

In the automatic speech recognition system described above, the sound source localization module is configured to execute a process comprising: performing a frequency analysis for the acoustic signals detected by the microphones to extract harmonic structures; acquiring an intensity difference and a phase

difference for the harmonic structures extracted through the plurality of microphones; acquiring belief factors for a sound direction based on the intensity difference and the phase difference, respectively; and determining a most probable sound direction.

[0009]

In the automatic speech recognition system described above, it may be preferable that the sound source separation module employs an active direction-pass filter so as to separate speeches, the filter being configured to execute a process comprising: separating speeches of a narrower directional band when a sound direction, which is localized by the sound source localization module, lies close to a front, which is defined by an arrangement of the plurality of microphones; and separating speeches of a wider directional band when the sound direction lies apart from the front.

[0010]

In the automatic speech recognition system described above, it may be preferable that the acoustic model composition module is configured to compose an acoustic model for the sound direction by applying weighted linear summation to the direction-dependent acoustic models in the acoustic model memory, and weights introduced into the linear summation are determined by training.

[0011]

In the automatic speech recognition system described above, it may be preferable that the system further comprises a speaker identification module for identifying the speaker, wherein the acoustic model memory possesses the direction-dependent acoustic models for respective speakers, and wherein the acoustic model composition module is configured to execute a process comprising:

based on direction-dependent acoustic models of a speaker who is identified by the speaker identifying module and a sound direction localized by the sound source localization module, obtaining an acoustic model for the sound direction based on the direction-dependent acoustic models in the acoustic model memory; and storing the acoustic model in the acoustic model memory.

[0012]

It may be preferable, but not necessarily, that the automatic speech recognition system further comprises a masking module. The masking module conducts comparison between patterns prepared in advance with the features extracted by the feature extractor or the speech signals separated by the sound source separation module so as to identify a domain, a frequency domain and sub-band, for example, in which a difference with respect to the patterns is greater than a predetermined threshold. The masking module sends an index indicating that reliability in terms of feature is low for the identified domain, to the speech recognition module.

[0013]

The automatic speech recognition system of the present invention, which identifies the sound direction of the acoustic signals generated in an arbitrary direction and carries out speech recognition using the acoustic model appropriate for the sound direction, is able to increase speech recognition rate.

[Preferred Embodiments]

[0014]

First Embodiment

Detailed description is given of an embodiment of the present invention with reference to the appended drawings. FIG.1 is a block diagram showing an automatic

speech recognition system according to a first embodiment of the present invention.

As shown in FIG.1, an automatic speech recognition system 1 according to the first embodiment includes two microphones M_R and M_L, a sound source localization module 10, a sound source separation module 20, an acoustic model memory 49, an acoustic model composition module 40, a feature extractor 30 and a speech recognition module 50. The module 10 localizes a position of a speaker (sound source) from acoustic signals detected by the microphones M_R and M_L. The module 20 separates acoustic signals originating from a sound source at a particular direction based on the direction of the sound source localized by the module 10 and spectrums obtained by the module 10. The memory 49 stores acoustic models adjusted to a plurality of directions. The module 40 composes an acoustic model adjusted to a sound direction, based on the sound direction which is localized by the module 10 and the acoustic models stored in the module 49. The extractor 30 extracts features of acoustic signals based on a spectrum of the specified sound source, which is separated by the module 20. The module 50 performs speech recognition based on the acoustic model composed by the module 40 and the features of the acoustic signals extracted by the extractor 30. Among these modules, the module 20 is not mandatory but adopted as the case may be.

The invention, in which the module 50 performs speech recognition with the acoustic model that is generated by the module 40 and adjusted to the sound direction, is able to realize a high recognition rate.

[0015]

Next, description is given of the microphones M_R and M_L, the sound source localization module 10, the sound source separation module 20, the feature extractor 30, the acoustic model composition module 40 and the speech recognition module

50, respectively where the microphones M_R and M_L, the module 10, the module 20, the extractor 30, the module 40 and the module 50 are elements configuring the automatic speech recognition system according to the embodiments of the present invention.

[0016]

(Microphones M_R and M_L)

The microphones M_R and M_L are each a typical type of microphone, which detects sounds and generates electric signals (acoustic signals). The number of microphones is not limited to two as is exemplarily shown in this embodiment, but it is possible to select any number, for example three or four, as long as it is plural. The microphones M_R and M_L are, for example, installed in the ears of a robot RB, which is a moving object.

A typical front of the automatic speech recognition system 1 is defined by an arrangement of the microphones M_R and M_L. It is described that a direction resulting from a sum of vectors, each being oriented to a sound direction collected by the microphones M_R and M_L, will coincide with the front of the automatic speech recognition system 1. As shown in FIG.1, when the microphones M_R and M_L are installed on left and right sides of a head of the robot RB, a front of the robot RB will coincide with the front of the automatic speech recognition system 1.

[0017]

(Sound source localization module 10)

FIG.2 is a block diagram showing an example of a sound source localization module. FIG.3 and FIG.4 are schematic diagrams each describing operation of a

sound source localization module.

The sound source localization module 10 localizes a direction of sound source for each of speakers HMn (HM1 and HM2 in FIG.3, for example) based on two kinds of acoustic signals received from the two microphones M_R and M_L. There are some methods for localizing a sound source, such as: a method for utilizing a phase difference between acoustic signals entered into the microphones M_R and M_L, a method for estimating with using head related transfer function of a robot RB and a method for establishing a correlation between signals entered through the right and left microphones M_R and M_L. Each of the methods described above has been improved in various ways so as to increase accuracy. Description is given here of a method as an example, with which the inventors of the present invention have succeeded in attaining improvement.

[0018]

As shown in FIG.2, the sound source localization module 10 includes a frequency analysis module 11, a peak extractor 12, a harmonic structure extractor 13, an IPD calculator 14, an IID calculator 15, a hypothesis 16 by auditory epipolar geometry, a belief factor calculator 17 and a belief factor integrator 18.

Each of these portions will be described with reference to FIG.3 and FIG.4. A situation where the speakers HM1 and HM2 simultaneously start speaking to the robot RB is assumed in the following description.

[0019]

(Frequency analysis module 11)

The frequency analysis module 11 cuts out a signal section having a microscopic time length Δt from right and left acoustic signals CR1 and CL1, which

are detected by the right and left microphones M_R and M_L installed in the robot RB, performing a frequency analysis for each of left and right channels with Fast Fourier Transform (FFT).

Results obtained from the acoustic signals CR1, which are received from the right microphone M_R, are designated as a spectrum CR2. Similarly, results obtained from the acoustic signals CL1, which are received from the left microphone M_L, are designated as a spectrum CL2.

It may be alternatively possible to adopt other methods for frequency analysis, such as a band pass filter.

[0020]

(Peak extractor 12)

The peak extractor 12 extracts consecutive peaks from the spectrums CR2 and CL2 for the right and left channels, respectively. One method is to directly extract local peaks of a spectrum. The other one is to use a method based on spectral subtraction method (See S. F. Boll, A spectral subtraction algorithm for suppression of acoustic noise in speech, Proceedings of 1979 International conference on Acoustics, Speech, and signal Processing (ICASSP-79)). The latter method extracts peaks from a spectrum, subtracts the extracted peaks from the spectrum, and generating a residual spectrum. A process for extracting peaks will be repeated until no peaks are found in the residual spectrum.

When extraction of peaks is carried out for the spectrums CR2 and CL2, only sub-band signals forming peaks such as peak spectrums CR3 and CL3 are extracted.

[0021]

(Harmonic structure extractor 13)

The harmonic structure extractor 13 generates a group, which contains peaks having a particular harmonic structure, for each of the right and left channels, according to harmonic structure which a sound source possesses. Taking a human voice, for example, a voice of a specified person is composed of sounds having fundamental frequencies and their harmonics. Because fundamental frequencies slightly differ from person to person, it is possible to categorize voices of a plurality of persons into groups according to difference in the frequencies. The peaks, which are categorized into a group according to harmonic structure, can be estimated as signals generated by a common sound source. If a plural number (j) of speakers is simultaneously speaking, for example, the same plural number (j) of harmonic structures is extracted.

[0022]

In FIG.3, peaks P1, P3 and P5 of the peak spectrum CR3 are categorized into one group of harmonic structure CR41. Peaks P2, P4 and P6 of the peak spectrum CR3 are categorized into one group of harmonic structure CR42. Similarly, peaks P1, P3 and P5 of the peak spectrum CL3 are categorized into one group of harmonic structure CL41. Peaks P2, P4 and P6 of the peak spectrum CL3 are also categorized into one group of harmonic structure CL42.

[0023]

(IPD calculator 14)

The IPD calculator 14 calculates an interaural phase difference (IPD) from spectrums of the harmonic structures CR41, CR42, CL41 and CL42 extracted by the

harmonic structure extractor 13.

The IPD calculator 14 selects a spectral sub-band corresponding to each harmonic of a peak frequency f_k contained in harmonic structure j (for example, the harmonic structure CR41) from both right and left channels (harmonic structures CR41 and CL41, for example), and calculates IPD $\Delta\phi(f_k)$ with an equation (1). The IPD$\Delta\phi(f_k)$ calculated from the harmonic structures CR41 and CL41 results in an interaural phase difference C51, as shown in FIG.4.

[0024]

$$\Delta\phi(f_k) = \arctan\left(\frac{\Im[S_r(f_k)]}{\Re[S_r(f_k)]}\right) - \arctan\left(\frac{\Im[S_l(f_k)]}{\Re[S_l(f_k)]}\right) \qquad (1)$$

where:

$\Delta\phi(f_k)$: IPD (interaural phase difference) for f_k

$J[S_r(f_k)]$: an imaginary part of spectrum for a peak f_k of right input signal

$R[S_r(f_k)]$: a real part of spectrum for a peak f_k of right input signal

$J[S_l(f_k)]$: an imaginary part of spectrum for a peak f_k of left input signal

$R[S_l(f_k)]$: a real part of spectrum for a peak f_k of left input signal

[0025]

(IID calculator 15)

The IID calculator 15 calculates a difference in sound pressure between sounds received from the right and left microphones M_R and M_L (interaural intensity difference) for each harmonic belonging to each harmonic structure.

The IID calculator 15 selects a spectral subband, which corresponds to a harmonic having a peak frequency f_k lying in a harmonic structure j (harmonic

structures CR41 and CL41, for example), from both right and left channels (harmonic structures CR41 and CL41, for example), and calculates an IID$\Delta\rho(f_k)$ with an equation (2). The IID$\Delta\rho(f_k)$ calculated from the harmonic structures CR41 and CL41 results in an interaural intensity difference C61 as shown in FIG.4, for example.

[0026]

$$\Delta\rho(f_k) = p_r(f_k) - p_l(f_k) \qquad (2)$$

where:

$\Delta\rho(f_k)$: IID (interaural intensity difference) for f_k

$Pr(f_k)$: power for peak f_k of a right input signal

$P_l(f_k)$: power for peak f_k of a left input signal

$Pr(f_k) = 10 \log_{10}(J[S_r(f_k)]^2 + R[S_r(f_k)]^2)$

$P_l(f_k) = 10 \log_{10}(J[S_l(f_k)]^2 + R[S_l(f_k)]^2)$

[0027]

(Hypothesis 16 by auditory epipolar geometry)

Let's see FIG.5, in which a head portion of the robot RB, which is modeled by a sphere, is viewed from upward. The hypothesis 16 by auditory epipolar geometry represents data of phase difference, which is estimated based on a time difference resulting from a difference in distance with respect to a sound source S between the microphones M_R and M_L, which are installed in both ears of the robot RB.

According to the auditory epipolar geometry, a phase difference $\Delta\phi$ is

obtained with an equation (3). It is assumed here that the sphere is representative of the shape of the head.

[0028]

$$\Delta\phi = \frac{2\pi f}{v} \times r(\theta + \sin\theta) \qquad (3)$$

[0029]

where $\Delta\phi$ represents an interaural phase difference (IPD), v is a sound velocity, f is a frequency, r is a value depending from an interaural distance 2r and θ represents a direction of a sound source.

The relationship between a phase difference $\Delta\phi$ and a frequency f of acoustic signals, which come from directions of each sound source, is obtained with the equation (3) and shown in FIG.6.

[0030]

(Belief factor calculator 17)

The belief factor calculator 17 calculates a belief factor for IPD and IID, respectively.

Description is first given of "IPD belief factor". An IPD belief factor is obtained as a function of θ so as to indicate which direction a harmonic component f_k is likely to come from, which is included in a harmonic structure j (harmonic structure CR41 or CL41, for example). The IPD is fitted into a probability function.

First, a hypothetical IPD (estimated value) for f_k is calculated with an equation (4).

[0031]

$$\Delta \phi_h(\theta, f_k) = \frac{2\pi f_k}{v} \times r(\theta + \sin\theta) \quad (4)$$

[0032]

$\Delta \phi_h(\theta, f_k)$ represents a hypothetical IPD (estimated value) with respect to a sound source lying in a direction θ for a kth harmonic component f_k in a harmonic structure. Thirty-seven hypothetical IPD's are, for example, calculated while a direction θ of a sound source is varied over a range from - 90 deg. to +90 deg. at intervals of 5 deg.. It may be alternatively possible to calculate at finer or rougher angle intervals.

Next, a difference between $\Delta \phi_h(\theta, f_k)$ and $\Delta \phi(f_k)$ is calculated with an equation (5) and a summation is obtained for all the peak frequencies f_k. This difference, which represents a distance between a hypothesis and an input, tends to take a smaller value if θ lies closer to a direction of a speaker but take a larger value if θ lies remoter from the direction of the speaker.

[0033]

$$d(\theta) = \frac{1}{K} \sum_{k=0}^{K-1} \frac{(\Delta \phi_h(\theta, f_k) - \Delta \phi(f_k))^2}{f_k} \quad (5)$$

[0034]

where $\Delta \phi(f_k)$ is an IPD of a harmonic component f_k contained in a harmonic structure and K represents number of harmonic components contained in the harmonic structure.

A belief factor $B_{IPD}(\theta)$ is obtained by substituting the resulting $d(\theta)$ into a probability density function, the following equation (6).

[0035]

$$B_{IPD}(\theta) = \int_{-\infty}^{X(\theta)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \qquad (6)$$

where $X(\theta) = (d(\theta) - m) / (\sqrt{(s/n)})$, m is a mean of $d(\theta)$, s is a variance of $d(\theta)$ and n is a number of hypothetical IPD's (37 in this embodiment).

[0036]

Description is given of "IID belief factor". An IID belief factor is obtained in the following manner. A summation of intensity differences of harmonic components included in a harmonic structure j is calculated with an equation (7).

[0037]

$$S = \sum_{k=0}^{K-1} \Delta \rho (f_k) \qquad (7)$$

[0038]

where K represents number of harmonics included in a harmonic structure, $\Delta\rho(f_k)$ is an IID calculated by the IID calculator 15.

Introducing Table 1, a likelihood to be right, center or left associated with a sound direction is transformed into a belief factor. In this connection, Table 1 shows empirical values.

When a hypothetical sound direction θ is equal to 40 deg. and an intensity difference S has a positive sign, for example, a belief factor $B_{IID}(\theta)$ is regarded as 0.35 according to the left-upper box of Table 1.

[0039]

[Table 1]

θ		90° ～30°	30° ～－30°	－30° ～－90°
S	＋	0．35	0．5	0．65
	－	0．65	0．5	0．35

[0040]

(Belief factor integrator 18)

The belief factor integrator 18 integrates an IPD belief factor $B_{IPD}(\theta)$ and an IID belief factor $B_{IID}(\theta)$ based on Dempster - Shafer theory with an equation (8), calculating an integrated belief factor $B_{IPD+IID}(\theta)$. A sound direction θ which provides a largest $B_{IPD+IID}(\theta)$ is considered to coincide with a direction of a speaker.

[0041]

$$B_{IPD+IID}(\theta)=1-(1-B_{IPD}(\theta))(1-B_{IID}(\theta)) \qquad (8)$$

[0042]

It may be alternatively possible to use a hypothesis by head related transfer function or a hypothesis by scattering theory instead of the hypothesis by auditory epipolar geometry.

(Hypothesis by head related transfer function)

A hypothesis by head related transfer function is a phase difference and an intensity difference for sounds detected by microphones M_R and M_L, which are obtained from impulses generated in a surrounding environment of a robot.

The hypothesis by head related transfer function is obtained in the following manner. The microphones M_R and M_L detect impulses, which are sent at appropriate

intervals (5 deg., for example) over a range of -90 deg. to 90 deg.. A frequency analysis is conducted for each impulse so as to obtain a phase response and a magnitude response with respect to frequency f. A difference between phase responses and a difference between magnitude responses are calculated to provide a hypothesis by head related transfer function.

The hypothesis by head related transfer function, which is calculated as described above, results in IPD shown in FIG.7A and IID shown in FIG.7B.

When a head related transfer function is introduced, it is possible to obtain a relationship between IID and a frequency of a sound coming from a certain direction with regard to the IID, in addition to IPD. Therefore, a belief factor is calculated based on distance data d(θ), which has been generated for both IPD and IID. The method for generating hypothesis is the same for IPD and IID.

Different from the method for generating a hypothesis with auditory epipolar geometry, a hypothesis by head related transfer function establishes a relationship between frequency f and IPD for a signal, which is generated in each sound direction, by means of measurement in lieu of calculation. Thus, a distance data d(θ), which is a distance between a hypothesis and an input, is directly calculated from actual measurement values shown in FIGs.7A and 7B, respectively.

[0043]

(Hypothesis by scattering theory)

Scattering theory estimates both IPD and IID, by taking into account waves scattered by an object, which scatters sounds, a head of a robot, for example. It is assumed here that a head of a robot is a sphere having a radius "a". It is also assumed that coordinates representative of the center of the head are an origin of a polar

coordinate.

When r_0 is a position of a point sound source and r is an observation point, a potential due to a direct sound at the observation point is defined by an equation (9).

$$V^i = \frac{v}{2\pi R\ \mathrm{f}} e^{i\frac{2\pi R\ \mathrm{f}}{\mathrm{v}}} \qquad (9)$$

where:

f : frequency of point sound source

v : sound velocity

R : distance between a point sound source and an observation point

A potential due to direct and scattering sound is defined by an equation (10) while the observation point r lies on a surface of the head.

$$S(\theta,\mathrm{f}) = V^i + V^S$$
$$= -\left(\frac{v}{2\pi a\ \mathrm{f}}\right)^2 \sum_{n=0}^{\infty} (2n+1) P_n(\cos\theta) \frac{h_n^{(1)}\left(\frac{2\pi r_0}{\mathrm{v}}\mathrm{f}\right)}{h_n^{(1)\prime}\left(\frac{2\pi a}{\mathrm{v}}\mathrm{f}\right)} \qquad (10)$$

where

V^S : potential due to scattering sound

P_n : Legendre Function of the First Kind

$h_n^{(1)}$: Spherical Hunkel Function of the First Kind

When polar coordinates for M_R and M_L are (a, $\pi/2$, 0) and (a,- $\pi/2$, 0), respectively, potentials at these microphones are represented by equations (11) and (12), respectively.

$$S_L(\theta, f) = S(f, \frac{\pi}{2} - \theta) \qquad \cdots \quad (11)$$

$$S_R(\theta, f) = S(f, -\frac{\pi}{2} - \theta) \qquad \cdots \quad (12)$$

In this way, a phase difference IPD$\Delta\phi_s(\theta, f)$ and an intensity difference IID$\Delta\rho_s(\theta, f)$ are calculated by the following equations (14) and (15), respectively.

$$\Delta\ \phi_S(\theta, f) = \arg(S_L(\theta, f)) - \arg(S_R(\theta, f)) \qquad (13)$$

$$\Delta\ \rho_S(\theta, f) = 20\log_{10}\frac{|S_L(\theta, f)|}{|S_R(\theta, f)|} \qquad (14)$$

[0044]

Replacing $\Delta\phi_h(\theta, f_k)$ of the equation (4) with IPD$\Delta\phi_s(\theta, f)$ of the equation (13), a $B_{IPD}(\theta)$ is calculated in the same process as that for auditory epipolar geometry.

Namely, a difference between $\Delta\phi_s(\theta, f_k)$ and $\Delta\phi(f_k)$ is calculated and a sum $d(\theta)$ for all peaks f_k is then calculated, which is substituted into the probability density function shown in the equation (6) so as to obtain a belief factor $B_{IPD}(\theta)$.

[0045]

As for IID, $d(\theta)$ and $B_{IID}(\theta)$ are calculated in the similar method to that applied to IPD. More specifically speaking, $\Delta\phi$ is regarded as $\Delta\rho$ and $\Delta\phi_h(\theta, f_k)$ in the equation (4) is replaced with IPD$\Delta\rho_s(\theta, f_k)$ in the equation (14). Then, a difference between $\Delta\rho_s(\theta, f_k)$ and $\Delta\rho(f_k)$ is calculated and a sum $d(\theta)$ for all peaks f_k is then calculated, which is substituted into the probability density function shown in the equation (6) so as to obtain a belief factor $B_{IID}(\theta)$.

[0046]

(Sound source separation module 20)

The sound source separation module 20 separates acoustic (speech) signals for each speaker HMn according to information on each localized sound direction by the sound source localization module 10 and a spectrum (spectrum CR2, for example) calculated by the sound source localization module 10. Though there may be conventional methods applicable to separation of a sound source, such as beam forming, null forming, peak tracking, a directional microphone, Independent Component analysis (ICA) and the like, for example, description here is given of a method with an active direction-pass filter developed by the inventors of the present invention.

As a sound direction lies remoter from the front of a robot RB, it tends to be more difficult to expect accuracy for information on the sound direction, which is estimated through two microphones, in separating a sound source by utilizing the information on the sound direction. In order to solve this problem, this embodiment employs active control so that a pass range is narrower for a sound source lying in the front direction but wider for a sound source lying remote from the front direction, thereby increasing accuracy for separating a sound source.

[0047]

More specifically speaking, the sound source separation module 20 includes a pass range function 21, a subband selector 22 and an acoustic signal re-composition module 23, as shown in FIG.8.

[0048]

(Pass range function 21)

As shown in FIG.9, the pass range function 21 is a function of a sound direction and a pass range, which is in advance adjusted to have a greater pass range as a sound direction lies remoter from the front. The reason for this is that it is more difficult to expect accuracy for information on a sound direction as it lies remoter from the front (0 deg.).

[0049]

(Subband selector 22)

The subband selector 22 selects a sub-band, which is estimated to come from a particular direction, out of respective frequencies (called "sub-band") of each of the spectrums CR2 and CL2. As shown in FIG.10, the subband selector 22 calculates IPD$\Delta\phi(f_i)$ and IID$\Delta\rho(f_i)$ (see an interaural phase difference C52 and an interaural intensity difference C62 in FIG.10) for sub-bands of each spectrum according to the equations (1) and (2), based on the right and left spectrums CR2 and CL2, which are generated by the sound source localization module 10.

Determining a θ_j, which is obtained by the sound source localization module 10, to be a sound direction which should be extracted, the subband selector 22 refers to the pass range function 21 so as to obtain a pass range width $\delta(\theta_j)$

corresponding to the θ_j. The subband selector 22 calculates a maximum θ_h and a minimum θ_l of the pass range according to the obtained pass range width $\delta(\theta_s)$ with the following equation (15).

[0050]

$$\left.\begin{aligned}\theta_l &= \theta_j - \delta(\theta_j)\\ \theta_h &= \theta_j + \delta(\theta_j)\end{aligned}\right\} \quad \cdots (15)$$

[0051]

A pass range B is shown in FIG.11 in the form of a plan view, for example.

Next, estimation is conducted for IPD and IID corresponding to θ_l and θ_h. This estimation is carried out with a transfer function, which is prepared in advance by measurement or calculation. The transfer function is a function which correlates a frequency f and IPD as well as a frequency f and IID, respectively, with respect to a signal coming from a sound direction θ. As described above, epipolar geometry, a head related transfer function or scattering theory is applied to the transfer function. An estimated IPD is, for example, shown in FIG.10 as $\Delta\phi_l(f)$ and $\Delta\phi_h(f)$ in an interaural phase difference C53, and an estimated IID is, for example, shown in FIG.10 as $\Delta\rho_l(f)$ and $\Delta\rho_h(f)$ in an interaural intensity difference C63.

[0052]

Utilizing a transfer function of a robot RB, the subband selector 22 selects a sub-band for a sound direction θ_s according to a frequency f of the input spectrum. The subband selector 22 selects a sub-band based on IPD if the frequency f is lower than a threshold frequency f_{th}, or based on IID if the frequency f is higher than the threshold frequency f_{th}. The subband selector 22 selects a sub-band which satisfies a conditional equation (16).

[0053]

$$\left.\begin{array}{l} f < f_{th} : \Delta\phi_l(f_i) \leqq \Delta\phi(f_i) \leqq \Delta\phi_h(f_i) \\ f \geqq f_{th} : \Delta\rho_l(f_i) \leqq \Delta\rho(f_i) \leqq \Delta\rho_h(f_i) \end{array}\right\} \quad \cdots (16)$$

[0054]

where f_{th} represents a threshold frequency, based on which one of IPD and IID is selected as a criterion for filtering.

According to this conditional equation, a subband of frequency f_i (an area with diagonal lines), in which IPD lies between $\Delta\phi_l(f)$ and $\Delta\phi_h(f)$, is selected for frequencies lower than the threshold frequency f_{th} in the interaural phase difference C53 shown in FIG.10. In contrast, a subband (an area with diagonal lines), in which IID lies between $\Delta\rho_l(f)$ and $\Delta\rho_h(f)$, is selected for frequencies higher than the threshold frequency f_{th} in the interaural intensity difference C63 shown in FIG.10. A spectrum containing selected sub-bands in this way is referred to as "selected spectrum" in this specification.

[0055]

(Acoustic signal re-composition module 23)

An acoustic signal re-composition module 23 recomposes an acoustic signal by conducting an inverse Fourier transform, from a selected spectrum of right or left, which is selected by the subband selector 22, and obtains the acoustic signal (see a speech signal C7 shown in FIG. 10), generated by a sound source in a specified range.

[0056]

There is an alternative method, which introduces a directional microphone for separating a sound source, instead of the sound source separation module 20 according to this embodiment described above. More specifically speaking, a

microphone with narrow directivity is installed on a robot RB. If the face of the robot is so controlled that the directional microphone is turned to a sound direction θ_j acquired by the sound source localization module 10, it is possible to collect only speeches coming from this direction.

In case of the method according to this directional microphone, if there is only a single directional microphone, a problem may arise that collection of speeches is limited to a single person. However, it may be possible to allow simultaneous collection of speeches of a plurality of people if a plurality of directional microphones are arranged at regular intervals of a predetermined angle so that it is possible to use speech signals sent by each directional microphone arranged for a sound direction.

[0057]

(Feature extractor 30)

The feature extractor 30 extracts features necessary for speech recognition from speech, which is separated by the sound source separation module 20. It is possible to use a linear spectrum, which results from frequency analysis of the speech, or Mel-Frequency Cepstrum Coefficient (MFCC) as features of speech. In this embodiment, description is given of an example with MFCC.

[0058]

As shown in FIG.12, the feature extractor 30 includes a log converter 31, a Mel frequency converter 32 and a cosine transformation module 33.

The log converter 31 converts an amplitude of selected spectrum, which is selected by the subband selector 22 into a logarithm, acquiring a linear log spectrum.

The Mel frequency converter 32 makes the linear log spectrum generated by

the log converter 31 pass through a bandpass filter of Mel frequency, acquiring a Mel frequency log spectrum, whose frequency is converted to Mel scale.

The cosine transformation module 33 carries out a cosine transformation for the Mel frequency log spectrum generated by the Mel frequency converter 32. A coefficient obtained by this cosine transformation results in MFCC.

[0059]

It may be possible to add a masking module, which gives an index (0 to 1), within or after the feature extractor 30 so that a spectrum subband is not considered to have reliable features when an input speech is deformed due to noise.

As a specified example, at the time of obtaining MFCC, firstly at the stage of a linear spectrum, comparison with a spectrum of a pattern prepared in advance is conducted, a domain of a spectrum, in which a difference from the pattern is greater than a predetermined threshold, is identified and this domain is identified as a subband which is subject to an influence of deformation of an input speech. Further, two MFCC where MFCC of increasing an identified subband by 0 and MFCC of increasing an identified subband by 1 are conducted, are acquired.

The acquired MFCCs are compared, an index ω having its value close to 0 is given to MFCC where a difference between the compared MFCCs is large and an index ω having its value close to 1 is given to MFCC where a difference between the compared MFCCs is small. Or an empirical threshold is determined and when MFCC is less than or equal to the empirical threshold, "1" is given to MFCC, and when MFCC is more than the empirical threshold, "0" is given to MFCC. This index ω is used at the time of the speech recognition.

[0060]

When a directional microphone is used for sound source separation, an

ordinary method of frequency analysis, such as an FFT and a bandpass filter or the like, is applied to a separated speech obtained by the directional microphone so as to obtain a spectrum.

[0061]

(Acoustic model composition module 40)

The acoustic model composition module 40 composes an acoustic model adjusted to each localized sound direction based on direction-dependent acoustic models, which are stored in the acoustic model memory 49.

As shown in FIG.13, the acoustic model composition module 40, which has an inverse cosine transformation module 41, a linear converter 42, an exponential converter 43, a parameter composition module 44, a log converter 45, a Mel frequency converter 46 and a cosine transformation module 47, composes an acoustic model for a direction θ by referring to direction-dependent acoustic models H(θ_n), which are stored in the acoustic model memory 49.

[0062]

(Acoustic model memory 49)

Direction dependent acoustic models H(θ_n), which are acoustic models adjusted to respective directions θ_n with respect to the front of a robot RB, are stored in the acoustic model memory 49. A direction-dependent acoustic model H(θ_n) is trained on MFCC of speech of a person uttered from a particular direction θ_n by way of Hidden Markov Model (HMM) with regard to plural persons. As shown in FIG.14, each direction-dependent acoustic model H(θ_n) employs a phoneme as a unit for recognition, storing a corresponding sub-model h(m, θ_n) of monophone for

each phoneme. In this connection, it may be possible that other units for recognition such as triphone and the like are adopted for generating a sub-model.

If there are seven sub-models at regular intervals of 30 deg. over a range -90 deg. to 90 deg. in terms of direction θ_n and each sub-model is composed of 40 pieces of monophone, the number of sub-models h(m, θ_n) results in $7 \times 40=280$.

A sub-model h(m, θ_n) has parameters such as number of states, a probability density distribution for each state and state transition probability. In this embodiment, the number of states for each phoneme is fixed to three: front (state 1), middle (state 2) and rear (state 3). The probability density distribution in this embodiment is fixed to a normal distribution. In this way, the acoustic model memory 49 according to this embodiment is trained on a state transition probability P and parameters of the normal distribution, namely a mean μ and a standard deviation σ.

[0063]

Description is given of steps for generating training data for a sub-model h(m, θ_n).

Speech signals, which include particular phonemes, are applied to a robot RB by a speaker (not shown) in a direction, for which an acoustic model is intended to generate. The feature extractor 30 converts the detected acoustic signals to MFCC, which the speech recognition module 50 to be described later recognizes. In this way, a probability for a recognized speech signal is obtained as a result for each phoneme. An acoustic model undergoes adaptive training, while a teaching signal indicative of a particular phoneme corresponding to a particular direction is given to the resulting probability. The acoustic model undergoes further training with phonemes and words of sufficient kinds (different speakers, for example) to learn a

sub-model.

When a speech for training is given, it may be possible to give another speech as noise in a direction different from that, in which generation of an acoustic model is intended. In this case, the speech separation module 20 separates only a speech, which lies in a direction intended for generating an acoustic model, and then the feature extractor 30 converts the speech to MFCC. In addition, if an acoustic model is intended for unspecified speakers, it may be possible for the acoustic model to be trained on their voices. In contrast, if an acoustic model is intended for specified speakers individually, it may be possible for the acoustic model to be trained on each speaker.

[0064]

The inverse cosine transformation module 41 to the exponential converter 43, restores an MFCC of probability density distribution to a linear spectrum. They carry out a reverse operation for a probability density distribution in contrast to the feature extractor 30.

[0065]

(Inverse cosine transformation module 41)

The inverse cosine transformation module 41 carries out inverse cosine transformation for MFCC, which is possessed by a direction-dependent acoustic model $H(\theta_n)$ stored in the acoustic model memory 49, generating a Mel log spectrum.

[0066]

(Linear converter 42)

The linear converter 42 converts frequencies of the Mel log spectrum, which is generated by the inverse cosine transformation module 41, to linear frequencies, generating a log spectrum.

[0067]

(Exponential converter 43)

The exponential converter 43 carries out an exponential conversion for the intensity of the log spectrum, which is generated by the linear converter 42, so as to generate a linear spectrum. The linear spectrum is obtained in the form of a probability density distribution of a mean μ and a standard deviation σ.

[0068]

(Parameter composition module 44)

As shown in FIG.15, the parameter composition module 44 multiplies each direction-dependent acoustic model $H(\theta_n)$ by a respective weights and makes a sum of the resulting products, composing an acoustic model $H(\theta_j)$ for a sound direction θ_j. The direction-dependent acoustic model $H(\theta_n)$ are each converted to a probability density distribution of linear spectrum by the inverse cosine transformation module 41, the linear converter 42 and the exponential converter 43, having parameters such as means μ_{1n}, μ_{2n}, μ_{3n}, standard deviations σ_{1n}, σ_{2n}, σ_{3n} and state transition probabilities P_{11n}, P_{12n}, P_{22n}, P_{23n}, P_{33n}. The module 44 composes an acoustic model for a sound direction θ_j by taking an inner product of these parameters and weights W_n, which are obtained beforehand by training and stored in the acoustic model memory 49. In other words, the module 44 composes an acoustic model for a sound direction θ_j by taking a linear summation of direction-dependent

acoustic models H(θ_n). In this connection, it will be described later how a weight W_n is introduced.

[0069]

When sub-models lying in H(θ_j) are composed, a mean μ_{1j} of the state 1 is calculated by an equation (17).

[0070]

$$\mu_{1j} = \frac{1}{\sum_{n=1}^{N} w_n} \sum_{n=1}^{N} w_n \mu_{1n} \quad \cdots \quad (17)$$

[0071]

Means μ_{2j} and μ_{3j} can be calculated similarly.

[0072]

For composition of a standard deviation σ_{1j} of the state 1, a covariance σ_{1j}^2 is calculated by an equation (18).

$$\sigma_{1j}^{\ 2} = \frac{1}{\sum_{n=1}^{N} w_n} \sum_{n=1}^{N} w_n \sigma_{1n}^{\ 2} \quad \cdots \quad (18)$$

[0073]

Standard deviations σ_{2j} and σ_{3j} can be obtained similarly. It is possible to calculate a probability density distribution with the obtained μ and ρ.

[0074]

Composition of a state transition probability P_{11j} for the state 1 is calculated by an equation (19).

[0075]

$$P_{11j} = \frac{1}{\sum_{n=1}^{N} w_n} \sum_{n=1}^{N} w_n P_{11n} \quad \cdots \quad (19)$$

[0076]

State transition probabilities P_{12j}, P_{22j}, P_{23j} and P_{33j} can be calculated similarly.

[0077]

Next, a probability density distribution is reconverted from a linear spectrum to MFCC by a log converter 45 through a cosine transformation module 47. Because the log converter 45, Mel frequency converter 46 and cosine transformation module 47 are similar to the log converter 31, Mel frequency converter 32 and cosine transformation 33, respectively, description in detail is not repeated.

[0078]

When a probability density distribution is composed in the form of a mixture normal distribution instead of a single normal distribution, a probability density distribution $f_{1n}(x)$ is calculated by an equation (20) instead of the calculation of the mean μ and standard deviation σ described above.

[0079]

$$f_{1j}(x) = \frac{1}{\sum_{n=1}^{N} w_n} \sum_{n=1}^{N} w_n f_{1n}(x) \quad \cdots \quad (20)$$

[0080]

Probability density distributions $f_{2j}(x)$ and $f_{3j}(x)$ can be calculated similarly.

[0081]

The parameter composition module 44 has the acoustic model described above stored in the acoustic model memory 49.

In this connection, the parameter composition module 44 carries out in real time such acoustic model composition while the automatic speech recognition system 1 is in operation.

[0082]

(Training of a weight W_{nm})

Next, a training method of the weight W_{nm} ("n" indicates direction and "m" indicates phoneme) is explained.

When a weight $W_{j/a/}$ of phoneme /a/ is trained with regard to the sound direction θ_j, firstly by setting the $W_{j/a/}$ of a value (vector) of the weight for an appropriate initial value, a trial of recognizing an appropriate phoneme line including /a/, for example, training data [/a//b//c/], is carried out with the composed acoustic model $H(\theta_j)$ using this $W_{j/a/}$. Specifically speaking, a phoneme of [/a//b//c/] is generated from a speaker located in the sound direction θ_j and is recognized. Here, although it may be preferable that training data is one phoneme /a/ itself, a phoneme line is used because the training with the phoneme line having a plurality of phonemes linked, results in a good training result.

FIG.17 exemplarily shows results of recognition. In the FIG.17, the result of recognition with the acoustic model $H(\theta_j)$, which is composed with the initial value $W_{j/a/}$, is shown in the first row, and results of recognition with the direction-dependent acoustic models $H(\theta_n)$ of a direction θ_n are shown in the second row or below. For example, it is shown that the recognition result with an

acoustic model $H(\theta_j)$ was a sequence of phonemes [/x//y//z/] and the recognition result with an acoustic model $H(\theta_{-90})$ was a sequence of phonemes [/x//y//c/].

[0083]

Seeing the first phoneme in FIG.17 after the first trial, when a corresponding phoneme is recognized from θ_j to θ_{90} of FIG.17, a weight W for a model corresponding to the direction is increased by Δd. Δd is set to be 0.05, for example, which is empirically determined. In contrast, when no corresponding phoneme is recognized for a direction, a weight W for a model corresponding to the direction is decreased by $k\,\Delta d/(n-k)$. In this way, a weight for a direction-dependent acoustic model having produced a correct answer is increased, but a weight for a direction-dependent acoustic model without a correct answer is decreased.

Since $H(\theta_n)$ corresponds with $H(\theta_{90})$ each other in the case of the example shown in FIG.17, corresponding weights $W_{n/a/}$ and $W_{90/a/}$ are increased by Δd, but other weights are decreased by $2\Delta d/(n-2)$.

[0084]

On the other hand, when there are no directions θ_n, in which a phoneme coinciding with the first phoneme is recognized and there is a dominant direction-dependent acoustic model $H(\theta_n)$ having a larger weight $W_{j/a/}$ relative to other directions, a weight is decreased for only this model $H(\theta_n)$ by Δd and weights for other models are increased by $k\Delta d/(n-k)$. Because the fact that any direction-dependent acoustic model $H(\theta_n)$ failed recognition implies that a current distribution of weights is inappropriate, a reduction in weight is implemented for the direction, in which the current weight works dominantly.

It is determined whether a weight is dominant or not is by checking whether the weight W_{nm} is larger than a predetermined threshold W_{th} (0.8 here, for example).

If there are no dominant direction-dependent acoustic models H(θ_n), only the maximum weight W_{nm} is decreased by Δd and other weights W_{nm} for other direction-dependent acoustic models H(θ_n) are increased by $\Delta d/(n-1)$.

And the trial described above is repeated with the updated weights $W_{j/a/}$.

[0085]

When the recognition of the acoustic model H(θ_j) results in a correct answer "/a/", the repetition is stopped, and recognition and training is moved to the next phoneme "/b/".

When a given number of trials (0.5/ Δd times, for example) does not allow the recognition result of an acoustic model H(θ_j) to be a correct answer and recognition of /a/ is not successful for example, the trial is moved to training of a next phoneme /b/. Weight $W_{j/a/}$ is updated by the same value as the distribution of weight $W_{j/b/}$ for a phoneme (phoneme /b/, for example), which is successfully recognized at last.

[0086]

The weights W obtained by training described above are stored in the acoustic model memory 49.

[0087]

(Speech recognition module 50)

Using an acoustic model H(θ_j) composed for a sound direction θ_j, the speech recognition module 50 recognizes speech of each speaker HMn separated and generates character information. Subsequently, the module 50 recognizes the speech by referring to the word dictionary to provide results of recognition. Since this method of speech recognition is based on an ordinary technique with Hidden Markov

Model, description in detail would be omitted.

When a masking module, which adds an index ω indicating a belief factor of each sub-band of MFCC, is disposed inside or after the feature extractor 30, the speech recognition module 50 carries out the following process.

In the recognition method using the general Hidden Markov Model, an output probability of state S within an acoustic model for a feature x (x represents a feature vector and uses MFCC in this embodiment) extracted by the feature extractor 30 is represented by f(x|S). If the index ω is given by the feature extractor 30, the speech recognition module 50 calculates a reliable component X_r of x and the output probability is obtained by an equation (21).

$$f(x_r \mid S) = \sum_{l=1}^{L} P_l f(x_r \mid l, S) \qquad \cdots \quad (21)$$

P_l : mixture distribution factor

L : number of distribution included in a condition

$X = X_r + X_u$

X_u : unreliable component of X

X_r : reliable component of X

$X_r(i) = \omega(i) \times X(i)$

i : dimension of MFCC

Using the obtained output probability and state transition probability, the module 50 performs recognition in the same manner as that of general Hidden Markov Model.

[0088]

Description is given of operation carried out by an automatic speech recognition system 1 configured as described above.

As shown in FIG.1, speeches of a plurality of speakers HMn (see FIG.3) are inputted into microphones M_R and M_L of a robot RB.

Sound directions of acoustic signals detected by the microphones M_R and M_L are localized by a sound source localization module 10. As described above, the module 10 calculates a belief factor with hypothesis by auditory epipolar geometry after conducting frequency analysis, peak extraction, extraction of harmonic structure and calculation of IPD and IID. Integrating the belief factor of IPD and IID, the module 10 subsequently regards the most probable θ_j as a sound direction (see FIG.2).

[0089]

Next, a sound source separation module 20 separates a sound corresponding to a sound direction θ_j. Sound separation is carried out in the following manner. First, the module 20 obtains upper limits $\Delta\phi_h(f)$ and $\Delta\rho_h(f)$, and lower limits $\Delta\phi_l(f)$ and $\Delta\rho_l(f)$ for IPD and IID for a sound direction θ_j with a pass range function. The module 20 selects sub-bands (selected spectrum) which are estimated to be a spectrum for the sound direction θ_j by introducing the equation (16) described above and these upper limits and lower limits. Subsequently, the module 20 converts the spectrum of the selected sub-bands by reverse FFT, thereby, transforming the spectrum into speech signals.

[0090]

A feature extractor 30 converts the selected spectrum separated by the sound

source separation module 20 into MFCC by a log converter 31, a Mel frequency converter 32 and a cosine transformation module 33.

[0091]

On the other hand, an acoustic model composition module 40 composes an acoustic model, which is considered appropriate for a sound direction θ_j, by receiving a direction-dependent acoustic model $H(\theta_n)$ stored in an acoustic model memory 49 and a sound direction θ_j localized by the sound source localization module 10.

The acoustic model composition module 40, which has an inverse cosine transformation module 41, a linear converter 42 and an exponential converter 43, converts the direction-dependent acoustic model $H(\theta_n)$ into a linear spectrum. A parameter composition module 44 composes an acoustic model $H(\theta_j)$ for a sound direction θ_j by taking an inner product of a direction-dependent acoustic model $H(\theta_n)$ and a weight W_j for a sound direction θ_j, which the module 44 reads out from the acoustic model memory 49. The module 40, which has a log converter 45, a Mel frequency converter 46 and a cosine transformation module 47, converts this acoustic model $H(\theta_j)$ in the form of a linear spectrum to an acoustic model $H(\theta_j)$ in the form of MFCC.

[0092]

Next, a speech recognition module 50 carries out speech recognition with Hidden Markov Model, by using the acoustic model $H(\theta_j)$ composed by the acoustic model composition module 40.

[0093]

Table 2 shows an example resulting from the performance of the speech recognition described above.

[0094]

[Table 2]

	Conventional method							This invention
Direction of acoustic model	-90°	-60°	-30°	0	30°	60°	90°	40°
Recognition rate of isolated word	20%	20%	38%	42%	60%	59%	50%	78%

[0095]

As shown in Table 2, when direction-dependent acoustic models were prepared for a range of -90 deg. to 90 deg. at regular intervals of 30 deg. and speech recognition was carried out for isolated words with each acoustic model in a direction of 40 deg. (conventional method), the best recognition rate was 60%, which was obtained by a direction-dependent acoustic model for a direction of 30 deg.. In contrast, when recognition of isolated words with an acoustic model for a direction of 40 deg., which was composed with a method according to this embodiment was made, resulted in attaining high recognition rate of 78%. Because it is possible for an automatic speech recognition system 1 according to this embodiment to compose an appropriate acoustic model adjusted for the direction on all such occasions even when speech is uttered from an arbitrary direction, high recognition rate can be realized. In addition, it is possible for the system 1, which is able to recognize speech uttered in an arbitrary direction, to implement speech recognition with high recognition rate while speech recognition from a moving sound source is made or a moving object (robot RB) itself is moving.

[0096]

Because it may be alternatively possible to store a small number of direction-dependent acoustic models, at intervals of 60 deg. or 30 deg. in terms of

sound direction, for example, it may be possible to reduce costs necessary for training of the acoustic models.

Because it is sufficient to carry out speech recognition for a single composed acoustic model, parallel processing is not required so as to carry out speech recognition for acoustic models representative of plural directions, which may lead to a reduction in calculation cost. Therefore, the automatic speech recognition system 1 according to this embodiment is appropriate for real-time processing and embedded use.

[0097]

The present invention is not limited to the first embodiment, which has been described so far, but it may be possible to implement alternatives such as modified embodiments described below.

[0098]

[Second embodiment]

A second embodiment of the present invention has a sound source localization module 110, which localizes a sound direction with a peak of correlation, instead of the sound source localization module 10 of the first embodiment. Because the second embodiment is similar to the first embodiment except for this difference, description would not be repeated for other modules.

(Sound source localization module 110)

As shown in FIG.18, the sound source localization module 110 includes a frame segmentation module 111, a correlation calculator 112, a peak extractor 113

and a direction estimator 114.

[0099]

(Frame segmentation module 111)

The frame segmentation module 111 segments acoustic signals, which have entered right and left microphones M_R and M_L, so as to generate segmental acoustic signals having a given time length, 100msec for example. Segmentation process is carried out at appropriate time intervals, 30msec for example.

[0100]

(Correlation calculator 112)

The correlation calculator 112 calculates correlation by an equation (22) for the acoustic signals of the right and left microphones M_R and M_L, which have been segmented by the frame segmentation module 111.

$$CC(T) = \int_0^T x_L(t)x_R(t+T)dt \qquad (22)$$

where:

CCT(T): correlation between $x_L(t)$ and $x_R(t)$

T: frame length

$x_L(t)$: input signal from the microphone L segmented by frame length T

$x_R(t)$: input signal from the microphone R segmented by frame length T

[0101]

(Peak extractor 113)

The peak extractor 113 extracts peaks from the resulting correlations. Peaks are selected in order of peak height while the number of the peaks is adjusted to the number of sound sources when it is known in advance. When the number of sound sources is not known, on the other hand, it may be possible to extract all peaks exceeding a predetermined threshold or to select a predetermined number of peaks in order of peak height.

[0102]

(Direction estimator 114)

Receiving the obtained peaks, the direction estimator 114 calculates a difference of distance "d" shown in FIG.19 by multiplying an arrival time difference D of acoustic signals entering the right and left microphones M_R and M_L by sound velocity "v". The direction estimator 114 then generates a sound direction θ_j by the following equation.

θ_j =arcsin (d/b)

[0103]

The sound source localization module 110, which introduces the correlation described above, is also able to estimate a sound direction θ_j. It is possible to increase a recognition rate with an acoustic model appropriate for the sound direction θ_j, which is composed by an acoustic model composition module 40 described above.

[0104]

[Third embodiment]

A third embodiment has an additional function that a sound source localization module performs speech recognition while it is checking if acoustic signals come from a same sound source, in addition to the first embodiment. Description would not be repeated for modules which are similar to those described in the first embodiment, bearing the same symbols.

An automatic speech recognition system 100 according to the third embodiment has an additional module, that is, a stream tracking module 60, in addition to the automatic speech recognition system 1 according to the first embodiment. Inputting a sound direction localized by a sound source localization module 10, the stream tracking module 60 tracks a sound source so that it checks if acoustic signals continue coming from the same sound source. If it succeeds in confirmation, the stream tracking module 60 sends the sound direction to a sound source separation module 20.

[0105]

As shown in FIG.21, the stream tracking module 60 has a sound direction history memory 61, a predictor 62 and a comparator 63.

[0106]

The sound direction history memory 61 stores time, a sound source direction recognized in the time and a pitch of the sound source (a fundamental frequency f_1 which a harmonic structure of the sound source possesses) of a sound source at this time, in the correlated form.

[0107]

The predictor 62 reads out the sound direction history of the sound source, which has being tracked so far, from the sound direction history memory 61.

Subsequently, the predictor 62 predicts a stream feature vector (θ_j, f_1) with a Kalman filter and the like from the history so far, which is made of a sound direction θ_j and a pitch f_1 at current time t1, sending the stream feature vector (θ_j, f_1) to the comparator 63.

[0108]

The comparator 63 receives from the sound source localization module 10 a sound direction θ_j of each speaker j and a pitch f_1 of the sound source at current time t1, which has been localized by the sound source localization module 10. The comparator 63 compares a predicted stream feature vector (θ_j, f_1), which is inputted by the predictor 62, and a stream feature vector (θ_j, f_1) resulting from a sound direction and a pitch, which are localized by the sound source localization module 10. If a resulting difference (distance) is less than a predetermined threshold, the comparator 63 sends the sound direction θ_j to the sound source separation module. The comparator 63 also makes the stream feature vector (θ_j, f_1) store in the sound direction history memory 61.

If the difference (distance) is greater than the predetermined threshold, the comparator 63 does not send the localized sound direction θ_j to the sound source separation module 20, so that speech recognition is not carried out. In this connection, it may be alternatively possible for the comparator 63 to send data, which indicates whether or not a sound source can be tracked, to the sound source separation module 20 in addition to a sound direction θ_j.

It may be alternatively possible to use only a sound direction θ_j without using a pitch f_1 in performing prediction.

[0109]

In the automatic speech recognition system 100 having the stream tracking

module 60 described above, a sound direction is localized by the sound source localization module 10, and the sound direction and a pitch are inputted to the stream tracking module 60. In the stream tracking module 60, the predictor 62 reads out a sound direction history stored in the sound direction history memory 61, predicting a stream feature vector (θ_j, f_1) at a current time t1. The comparator 63 compares a stream feature vector (θ_j, f_1) which is predicted by the predictor 62 and a stream feature vector (θ_j, f_1) resulting from values, which are inputted by the sound source localization module 10. If the difference (distance) is less than a predetermined threshold, the comparator 63 sends a sound direction to the sound source separation module 20.

The sound source separation module 20 separates sound sources based on spectrum data, which is inputted by the sound source localization module 10, and sound direction θ_j data, which is outputted by the stream tracking module 60, in the similar manner as that of the first embodiment. A feature extractor 30, an acoustic model composition module 40 and a speech recognition module 50 carry out processes in the similar manner as that of the first embodiment.

[0110]

Because the automatic speech recognition system 100 according to this embodiment carries out speech recognition as a result of checking if a sound source can be tracked, it is able to keep carrying recognition for a speech uttered by the same sound source even if the sound source is moving, which will lead to a reduction in probability for false recognition. The automatic speech recognition system 100 is beneficial for a situation where there is a plurality of moving sound sources, which intersect each other.

In addition, the automatic speech recognition system 100, which not only

stores but also predicts sound directions, is able to reduce an amount of processing if searching for a sound source is limited to a certain area corresponding to a particular sound direction.

[0111]

While the embodiments of the present invention have been described, the present invention is not limited to these embodiments, but can be implemented with various changes and modifications.

One example is an automatic speech recognition system 1, which includes a camera, a well-known image recognition system and a speaker identification module, which recognizes a face of a speaker and identifies the speaker referring to its database. When the system 1 possesses the direction-dependent acoustic models for each speaker, it is possible to compose an acoustic model appropriate for each speaker, which enables higher recognition rate. It may be possible to adopt an alternative, which introduces speeches of speakers registered in advance in the form of vector by vector quantization (VQ). The system 1 compares the registered speeches and a speech in the form of vector, which the sound source separation module 20 separates, localizing the speaker by outputting the speaker having the smallest distance as a result.

[Brief Description of the Drawings]

[0112]

FIG.1 is a block diagram showing an automatic speech recognition system according to an embodiment of the present invention.

FIG.2 is a block diagram showing an example of a sound source localization module.

FIG.3 is a schematic diagram illustrating operation of a sound source localization module.

FIG.4 is a schematic diagram illustrating operation of a sound source localization module.

FIG.5 is a schematic diagram describing auditory epipolar geometry.

FIG.6 is a graph showing the relationship between phase difference $\Delta\phi$ and frequency f.

FIG.7 is a graph showing an example of a head related transfer function.

FIG.8 is a block diagram showing an example of a sound source separation module.

FIG.9 is a graph showing an example of a pass range function.

FIG.10 is a schematic diagram illustrating operation of a subband selector.

FIG.11 is a plan view showing an example of a pass range.

FIG.12 is a block diagram showing an example of a feature extractor.

FIG.13 is a block diagram showing an example of an acoustic model composition module.

FIG.14 is a table showing a unit for recognition and a sub-model of a direction-dependent acoustic model.

FIG.15 is a schematic diagram illustrating operation of a parameter

composition module.

FIG.16 is a graph showing an example of a weight W_n.

FIG.17 is a table showing a training method of a weight W.

FIG.18 is a block diagram showing an automatic speech recognition system according to a second embodiment of the present invention.

FIG.19 is a schematic diagram illustrating a difference in input distance of an acoustic signal.

FIG.20 is a block diagram showing an automatic speech recognition system according to a third embodiment of the present invention.

FIG.21 is a block diagram showing a stream tracking module.

FIG.22 is a graph showing a sound direction history.

[Explanation of Reference marks]

[0113]

Below, marks used in the accompanying drawings will be listed up for more of understanding and confirmation.

1: automatic speech recognition system, 10: sound source localization module, 11: frequency analysis module, 12: peak extractor, 13: harmonic structure extractor, 14: IPD calculator, 15: IID calculator, 16: hypothesis by auditory epipolar geometry, 17: belief factor calculator, 18: belief factor integrator, 20: sound source separation module, 21: pass range function, 22: subband selector, 23: acoustic signal recomposition module, 30: feature extractor, 40: acoustic model composition

module, 49: acoustic model memory, 50: speech recognition module, M: microphone.

ABSTRACT OF THE DISCLOSURE

[Abstract]

[Object] There is provided a speech recognition device capable of recognizing a speech with a high accuracy even when a speaker or mobile body having the speech recognition device is moving.

[Means for solution] The speech recognition device recognizes a speech of a particular speaker HMn from an acoustic signal detected by a plurality of microphones M and converts it into character information. The speech recognition device includes: a sound source localization module 10 for localizing the sound source direction θ_j of the speaker HMn based on the acoustic signal detected by a plurality of microphones M; a sound source separation module 20 for separating the speech signal of the speaker HMn from the acoustic signal according to the sound source direction θ_j; an acoustic model memory 49 storing a direction-dependent acoustic model $H(\theta_n)$ corresponding to intermittent plurality of directions; an acoustic model composition module 40 for obtaining an acoustic model of the sound source direction θ_j according to the direction-dependent acoustic model $H(\theta_n)$ of the acoustic model memory 49 and making the acoustic model memory 49 store the result; and a speech recognition module 50 for using the acoustic model composed by the acoustic model composition module 40 so as to perform speech recognition of the speech signal separated by the sound source separation module 20 and convert the speech into character information.

[Selected Drawing] FIG. 1